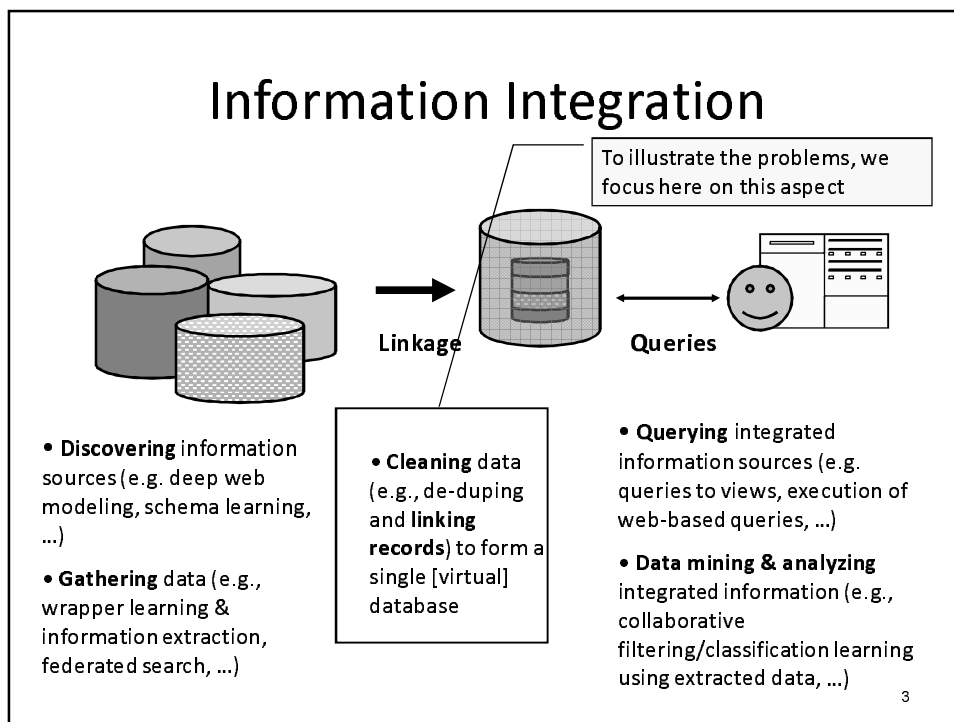


Data Integration

An Overview

What is Information Integration and
Why is it important

Some of the upcoming slides are from William
Cohen's tutorial on information integration
(WebDB 2005)



Automatic Linkage of Vital Records*

[Science 1959]

Computers can be used to extract "follow-up" statistics of families from files of routine records.

H. B. Newcombe, J. M. Kennedy, S. J. Axford, A. P. James

The term *record linkage* has been used to indicate the bringing together of two or more separately recorded pieces of information concerning a particular individual or family (*I*). Defined in this broad manner, it includes almost any use of a file of records to determine what has subsequently happened to people about whom one has some prior information.

***Record linkage*: bringing together of two or more separately recorded pieces of information concerning a particular individual or family (Dunn, 1946; Marshall, 1947).**

4

Motivations for Record Linkage c. 1959

The need for various follow-up studies such as might be carried out with the aid of record linkage have been discussed in detail elsewhere (1, 2), and there are numerous examples of important surveys which could be greatly extended in scope if existing record files were more readily linkable (3). Our

special interest in the techniques of record linkage relates to their possible use

(i) for keeping track of large groups of individuals who have been exposed to low levels of radiation, in order to determine the causes of their eventual deaths (see 4, chap. 8, para. 48; 5),

and (ii) for assessing the relative importance of repeated natural mutations on the one hand, and of fertility differentials on the other, in maintaining the frequency of genetic defects in human populations (see 4, chap. 6, para. 36c).

Our own studies (6) were started as part of a plan to look for possible differentials of family fertility in relation to the presence or absence of hereditary disease (through the use of vital records and a register of handicapped children).

Record linkage is motivated by certain problems faced by a small number of scientists doing data analysis for obscure reasons.

5

Information integration in 1959

- Many of the basic principles of modern integration work are recognizable.
- *Manual engineering* of distance features (e.g., last names as Soundex codes) that are then matched *probabilistically*.
 - $DB_1 + DB_2 + \text{Pr}(\text{matches}) + \text{elbowGrease} \rightarrow DB_{12}$
- Applied to records from *pairs* of datasets
 - “Smallest possible scale” for integration (one one dimension)
- Computationally *expensive*
 - Relative to ordinary database operations
- Narrowly used
 - Only for *scientists* in certain *narrow areas* (e.g., public health)
- ***How can this process be fully automated?***
- ***Why should we care?***

6

Ted Kennedy's "Airport Adventure" [2004]

Washington -- Sen. Edward "Ted" Kennedy said Thursday that he was stopped and questioned at airports on the East Coast five times in March because his name appeared on the government's secret "no-fly" list... Kennedy was stopped because the name "T. Kennedy" has been used as an alias by someone on the list of terrorist suspects.


"...privately they [FAA officials] acknowledged being embarrassed that it took the senator and his staff more than three weeks to get his name removed."

San Francisco Chronicle

Terror no-fly list singled out Kennedy Senator was stopped 5 times at airports

Sara Kehaulani Goo, Washington Post
Friday, August 20, 2004

Washington -- Sen. Edward "Ted" Kennedy said Thursday that he was stopped and questioned at airports on the East Coast five times in March because his name appeared on the government's secret "no-fly" list.



Federal air security officials said the initial error that led to scrutiny of the Massachusetts Democrat should not have happened even though they recognize that the no-fly list is imperfect. But privately they acknowledged being embarrassed that it took the senator and his staff more than three weeks to get his name removed.

A senior administration official, who spoke on condition he not be identified, said Kennedy was stopped because the name "T. Kennedy" has been used as an alias by someone on the list of terrorist suspects.

[Printable Version](#)
[Email This Article](#)

Florida Felon List [2000,2004]

USA TODAY Classifieds: Cars | Jobs | Dating

Home News Travel Money Sports Life Tech Weather Search

Nation

Posted 7/19/2004 7:21 PM

Fla. scraps flawed felon voting list

MIAMI (AP) — Florida elections officials said Saturday they will not use a disputed list that was designed to keep felons from voting, acknowledging a flaw that could have allowed convicted Hispanic felons to cast ballots in November.

The glitch in a state that President Bush won by just 537 votes could have been significant — because of the state's sizable Cuban population, Hispanics in Florida have tended to vote Republican... The list had about 28,000 Democrats and around 9,500 Republicans...

The purge of felons from voter rolls has been a thorny issue since the 2000 presidential election. A private company hired to identify ineligible voters before the election produced a list with scores of errors, and elections supervisors used it to remove voters without verifying its accuracy...

The new list ... contained few people identified as Hispanic; of the nearly 48,000 people on the list created by the Florida Department of Law Enforcement, only 61 were classified as Hispanics.

Gov. Bush said the mistake occurred because two databases that were merged to form the disputed list were incompatible. ... when voters register in Florida, they can identify themselves as Hispanic. But the potential felons database has no Hispanic category...

CNN.com International Edition
MEMBER SERVICES

SEARCH The Web CNN.com Search PowerRank

Home Page
World
U.S.
Weather
Business at cnnmoney.com
Sports at si.com
Politics
Law
Technology
Science & Space
Health
Entertainment
Travel
Education
Special Reports
Autos with comautos.com

Advertisement

SERVICES
Video
E-mail Newsletters
Your E-mail Alerts

U.S.
Report: FBI wasted millions on 'Virtual Case File'

Mueller says he'll decide on software

From Terry Frieden
CNN Washington Bureau
Thursday, February 3, 2005 Posted: 11:06 PM

WASHINGTON (CNN) -- FBI Director Robert Mueller promised a Senate panel late Thursday that he will decide within two months whether to scrap special computer software for FBI agents after a report sharply criticized the program.

Whatever his decision, Mueller told senators he believes FBI agents will have the software they need within one year.

Information dealing with such matters as violent crime, organized crime, fraud and other white-collar crime **may take days to be shared** throughout the law enforcement community, according to an FBI official.

The new software program was supposed to allow agents to pass along intelligence and criminal information in real time....

In a response contained in the inspector general's report, the FBI pointed to its Investigative Data Warehouse...that provides ... access to **47 sources of counterterrorism data**, including information from FBI files, other government agencies and open-source news feeds.

9

INFORMATION AWARENESS OFFICE
Scientia Est Potentia DARPA

Home
News
Programs
Solicitations

June 10, 2005 12:03pm ET The Early Show CBS Evening News 48 Hours

WAR ON TERROR Section Front
E-mail This Story Printable Version

Senate Targets DoD Spy Program

WASHINGTON, July 16, 2003

..cour
inform

(CBS) Without fanfare, senators debating defense spending for next year have proposed eliminating all money for the Pentagon's development of a vast computerized terrorism surveillance program that has raised privacy concerns.

10

Chinese Embassy Bombing [1999]

- May 7, 1999: NATO bombs the Chinese Embassy in Belgrade with five precision-guided bombs—sent to the wrong address—killing three.

“The Chinese embassy was mistaken for the intended target...located just 200 yards from the embassy. Reliance on an outdated map, aerial photos, and the extrapolation of the address of the federal directorate from number patterns on surrounding streets were cited ... as causing the tragic error...despite the **elaborate system of checks** built-into the targeting protocol, the coordinates did not trigger an alarm **because the three databases used in the process all had the old address.**” [US-China Policy Foundation summary of the investigation]

“BEIJING, June 17 — China today publicly rejected the U.S. explanation ... [and] said the U.S. report ‘does not hold water.’” [Washington Post]

“The Chinese embassy was clearly marked on **tourist maps** that are on sale internationally, including in the English language. ... Its address is listed in the **Belgrade telephone directory**... For the CIA to have made such an elementary blunder is simply not plausible.” [World Socialist Web Site]

“Many observers believe that the bombing was deliberate...it if you believe that the bombing was an accident, you already believe in the far-fetched” [disinfo.com, July 2002].

11

Information integration in 2005

- Apparently, we still have work to do.
 - Why is this problem so hard?
 - ➔ – **The airport adventure:** When can you tell if “*T. Kennedy*” the same person as “*Ted Kennedy*”? When can you **accept** an answer of “*I don’t know*”? What sorts of **information** can you use in deciding: structured data, text, **images**, ... ?
 - **The embassy bombing:** When are **multiple sources** that agree really useful? When have you looked at enough? What are the implications of looking at many sources?
 - **The felon list:** If you act on uncertain matches, what kind of **errors** will you make? will they cancel out, or accumulate?

12

Information integration in 2005

- It is hard to give Definitions: What do we really mean when we say “X is the same as Y”? does every user mean the same thing?
- Is “X is the same as Y” transitive?
- What conclusions follow from “X is the same as Y”?
 - Is it true that: Istanbul = Constantinople?
 - Does it follow that: The capital of Byzantium = Istanbul?

13

When *are* two entities are the same?



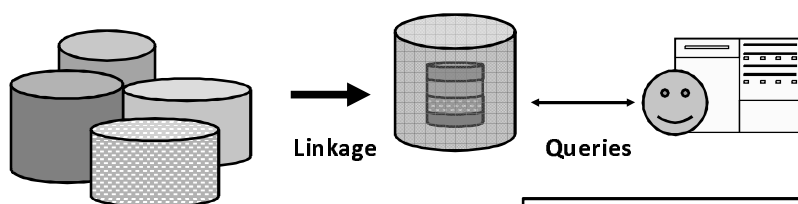
14

Information integration in 2005

- Apparently, we still have work to do.
- We fail to integrate information correctly
 - “Ted Kennedy (senator)” ≠ “T. Kennedy (terrorist)”
- Crucial decisions are affected by these errors
 - Who can/can't vote (felon list)
 - Where bombs are sent (Chinese embassy)
- Storing, linking, and analyzing information is a double-edged sword:
 - Loss of privacy and “fishing expeditions”

15

Information Integration: today and tomorrow



- **Discovering** information sources: based on standards and free-text metadata.

- Data providers will be even more numerous.

- **Gathering** data: will get cheaper and cheaper

- **Cleaning** data to form a single virtual database will be guided by a *user* or group of users, and by characteristics of *all* the data

- **Querying** integrated information sources may be done in radically different query models

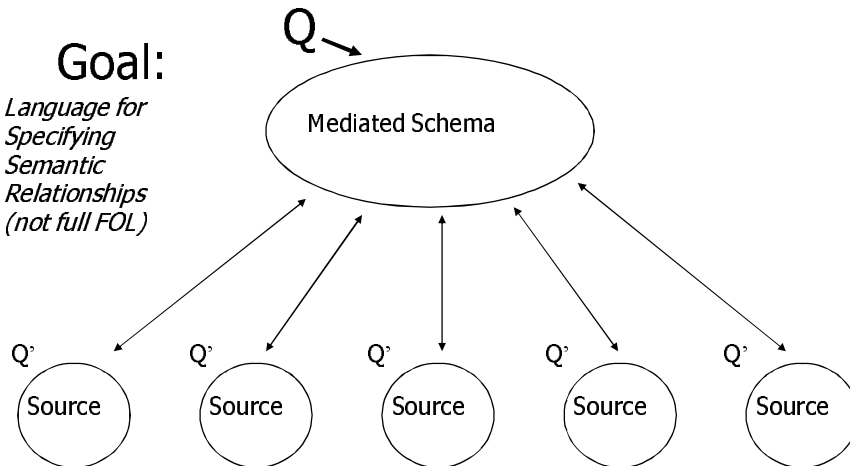
- **Data mining & analyzing** integrated information will be the norm, not the exception

16

Mediation Languages

Goal:

*Language for
Specifying
Semantic
Relationships
(not full FOL)*

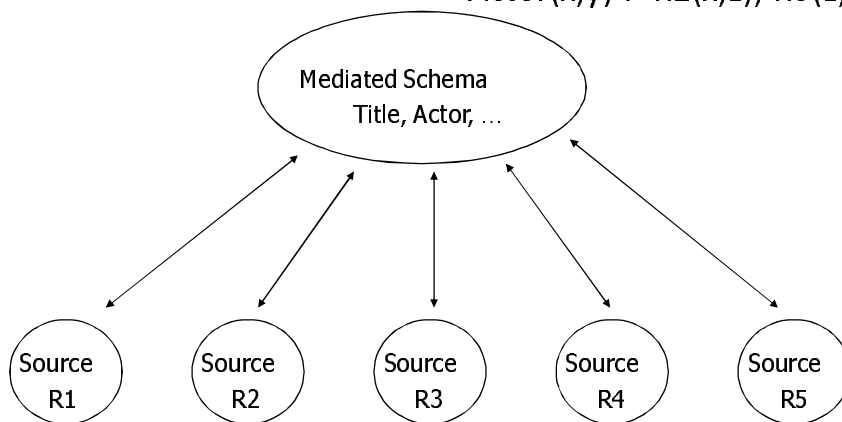


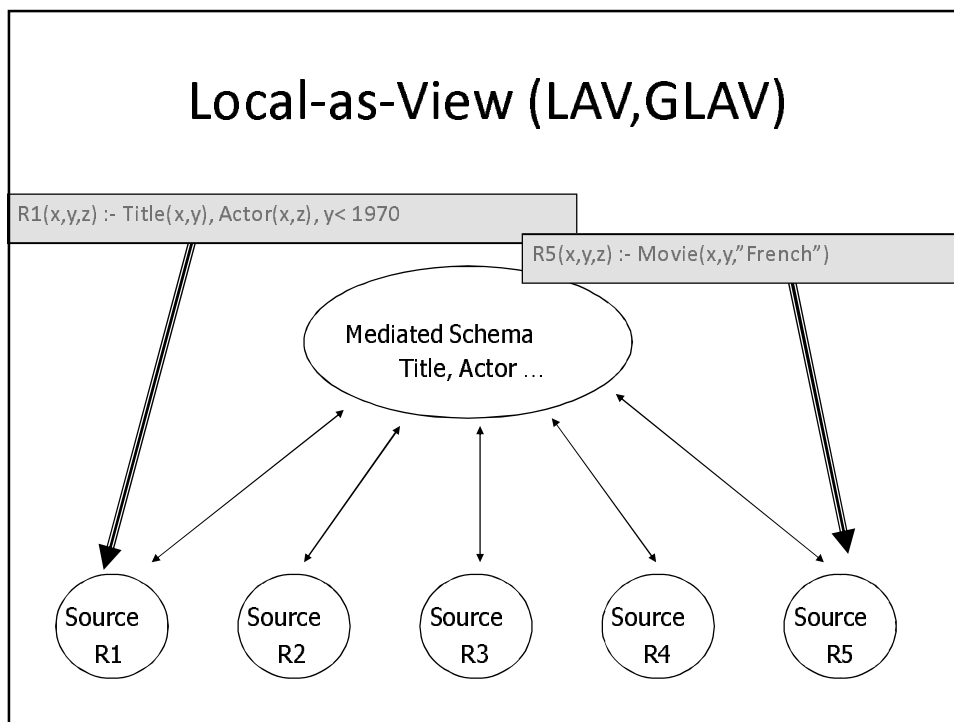
Assume: data at the sources is structure (or seems so).

Global-as-View (GAV)

$\text{Actor}(x,y) \text{ :- } R1(x,y,z)$

$\text{Actor}(x,y) \text{ :- } R2(x,z), R3(z,y)$





LAV vs. GAV

- What are the advantages of LAV?
- What are the advantages of GAV?
- How are queries over the entire data being answered in each approach?
 - GAV – Unfolding (easy)
 - LAV - Answering queries using views (NP-hard)

Queries in LAV

- Suppose that we have the following mapping rules:

ActingInfo(title, aname, year) \rightarrow Actor(aname, address)
 ActingInfo(title, aname, year) \rightarrow Movie(title, year, director)

- How does the data look like?
- We need to deal with incomplete information!
- How can we answer queries?

ActorInfo(n, a) \leftarrow Actor(n, a)
 Titles(t,y) \leftarrow Movie(t,y,d)

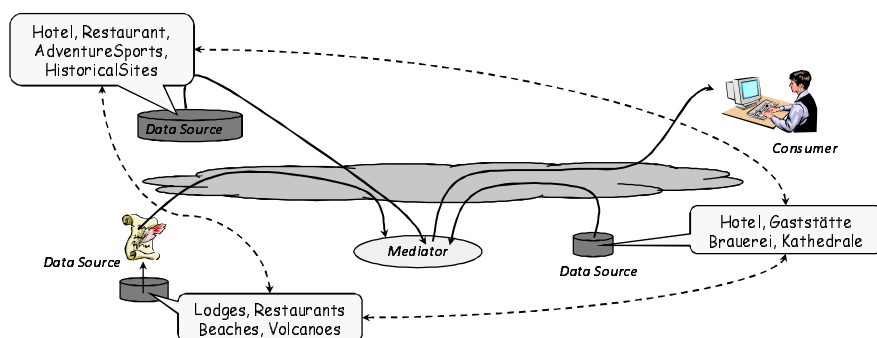
Dealing with Incomplete Information

- Given an incomplete database D' (i.e., there are predicates with null values), we consider all the possible completions \mathbf{D} of D'
- Given a query Q over D' , a certain answer A of Q is an answer that is given for any possible completion, i.e., for any database of \mathbf{D}
- We consider query answering as the set of all certain answers
- How do we deal with negation (e.g., not exists)?

Maximal Answers

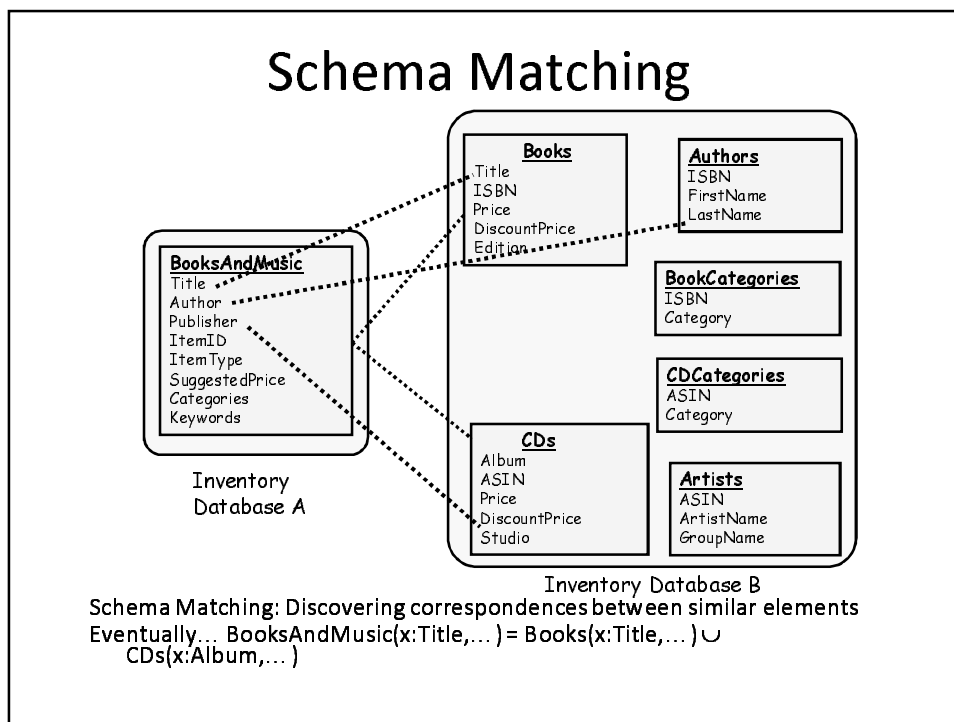
- One approach to deal with missing values is by computing maximal answers:
 - Full disjunction in the relational case
 - Different semantics of maximal matching in the case of matching graph queries to graph databases
 - In both cases, computation is intricate

Schema/Ontology Matching



Schema heterogeneity: a key roadblock for information integration

- Different data sources speak their own schema
- Mapping is key to *any* data sharing architecture



Typical Approaches

- Multiple sources of evidences in the schemas
 - Schema element names
 - BooksAndCDs/Categories ~ BookCategories/Category
 - Descriptions and documentation
 - ItemID: unique identifier for a book or a CD
 - ISBN: unique identifier for any book
 - Data types, data instances
 - DateTime ≠ Integer, In isolation, techniques are incomplete or brittle
 - addresses have similar formats
 - Schema structure
 - All books have similar attributes
 - Use domain knowledge

Combine multiple techniques to exploit all available evidence

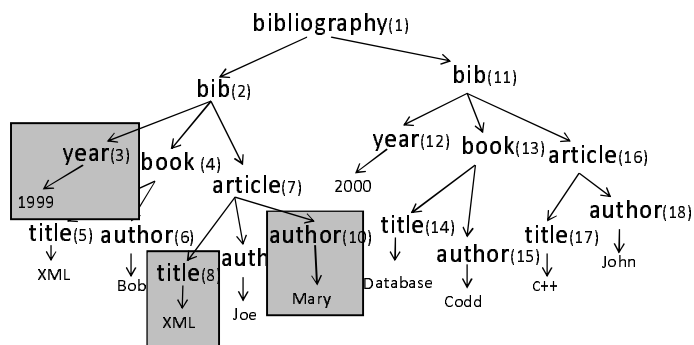
XML

- In XML there is no strict schema
- Integration is easier: you simply take XML from different sources and put them in a single repository
- Well, actually the main problem of linking related pieces of information remains!
- And, additional new problems emerge (to whom is it good?) 😊

Querying and Searching in XML

- Some challenges arise:
 - How to deal with variations in the structure of the XML?
 - How to deal with incomplete information?
 - How to find meaningful relationships among elements? An important example – keyword search.

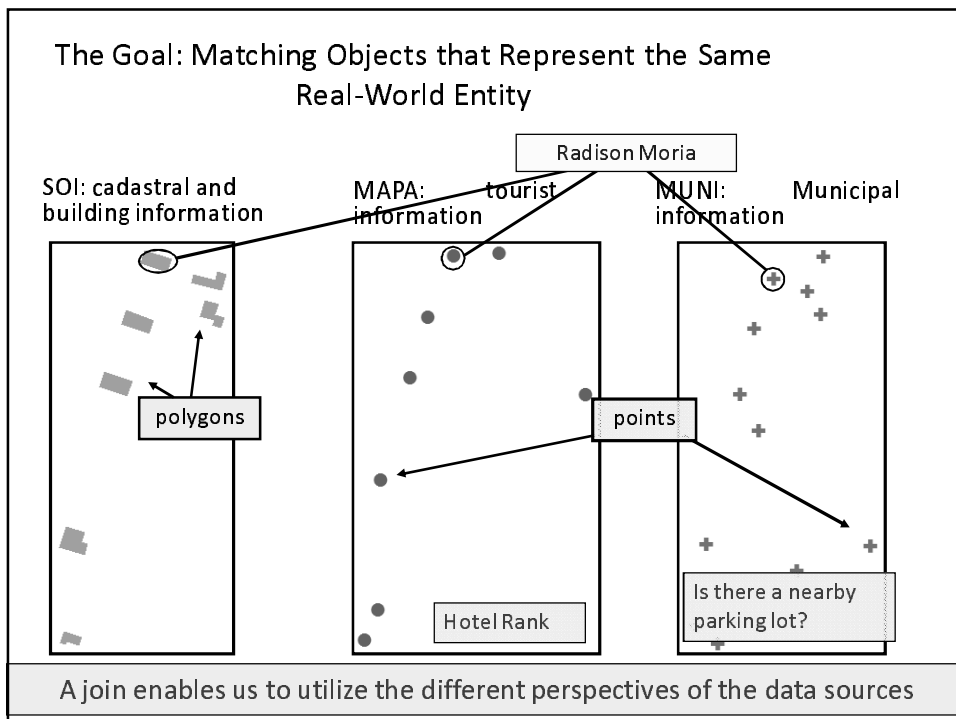
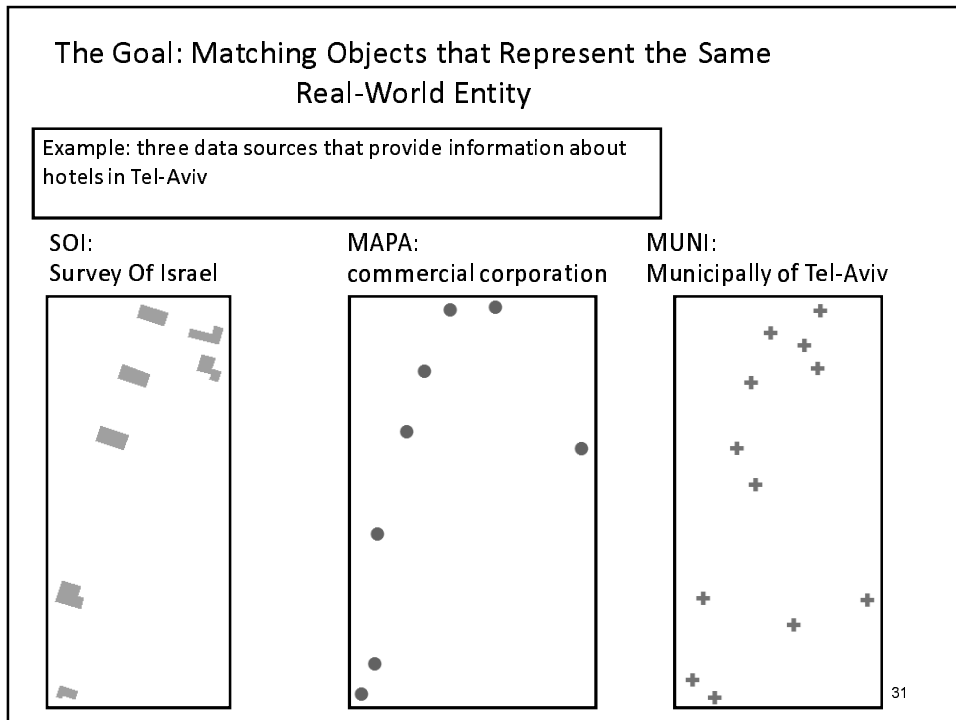
An example

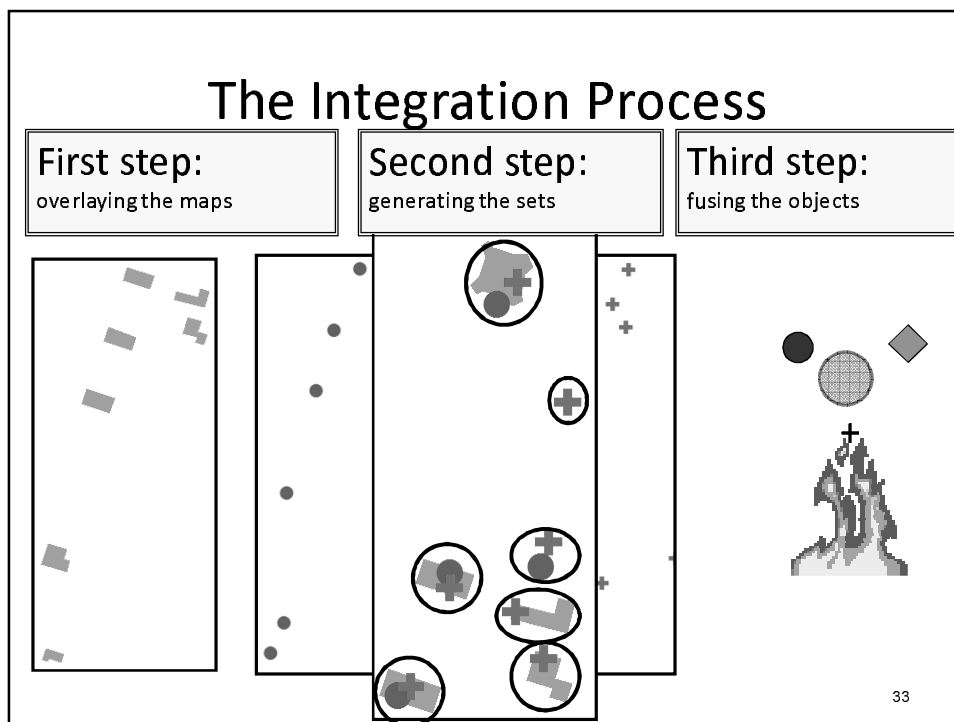


Query: What are the titles and years of the publications, of which Mary is an author?

Integration of Geographic Data

- The goal: Matching objects that represent the same real-world entity in different maps





Questions about Integration of Geographic Data

- How can we integrate efficiently and effectively geographical datasets?
- How does the existence of road networks affect the integration?
- Can a schema or ontology help us?

Using Locations for Matching Objects

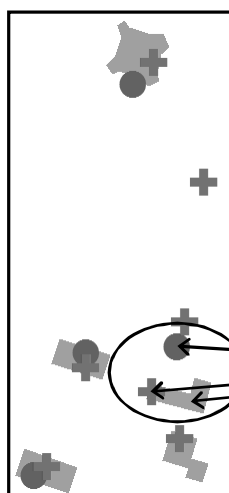
- There are no **global keys** to identify objects that should be joined
- Names cannot be used
 - Change often
 - May be missing
 - May be in different languages
- It seems that locations are **keys**:
 - Each spatial object contains location attributes
 - In a “perfect world,” two objects that represent the same entity have the same location

Global key = common identifier in the different sources



35

Locations are Inaccurate



- In real maps, locations are inaccurate
- The map on the left is an overlay of the three data sources about hotels in Tel-Aviv

For example, the Basel Hotel has three different locations, in the three data sources

36

Semantic Web

“Most of the Web's content today is designed for humans to read, not for computer programs to manipulate meaningfully.”

Berners-Lee, T, Hendler, J & Lassila, O ‘The semantic web’, *Scientific American*, May 2001

The Semantic Web

“For the semantic web to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning.”

Berners-Lee, T, Hendler, J & Lassila, O ‘The semantic web’, *Scientific American*, May 2001

The Semantic Web

- The main idea: Add semantics and reasoning instead of applying artificial intelligence
- Basic standards being developed: XML, XSchema, RDF, RDFS, OWL
- Is the Semantic Web the holly grail of integration?

Privacy

- How can we publish information and yet, guarantee that integration won't reveal sensitive data?